

ENHANCING ON-CHIP NETWORK PREDICTIONS WITH ADVANCED AI TECHNIQUES

Interim Report for Research Project Module 5E1

Lingyu Gong
Trinity College Dublin
gongl@tcd.ie

January 2024

This report is submitted in part fulfillment for the assessment required in 5E1 Research Project. I have read and understand the plagiarism provisions in the General Regulations of the University Calendar for the current year. These are found in Parts II and III at <http://www.tcd.ie/calendar>.

Supervisor
Dr. Libin Mathew

Abstract

With the growing complexity of mobile applications, the evolution of system-on-chip (SoC) technology has become increasingly critical, demanding higher integration, reduced latency, and improved energy efficiency. The progression of SoC from single-core to multi-core, and now to many-core systems, has unfortunately introduced performance bottlenecks. This challenge has steered the focus toward the development and exploration of Network-on-Chip (NoC) technology.

While several platforms for compiling and emulating NoCs exist, their scalability issues become apparent as the project size increases. The time-intensive nature of these general emulation platforms hinders the system's overall optimization. This limitation has prompted the innovative integration of Artificial Intelligence (AI) with NoC technology.

This project concentrates on harnessing AI to augment NoC technology, aiming to enhance the efficiency of mobile computing. By leveraging NoC for more efficient communication and reduced power consumption, the project addresses key challenges in mobile applications. Utilizing Booksim software, the project involves the configuration of NoC networks, followed by the development of AI algorithms. These algorithms are designed to optimize hardware efficiency and adapt to the dynamic demands of mobile computing.

The primary goal is to create scalable, efficient systems by synergizing AI's computational prowess with NoC's advanced communication capabilities. This approach not only addresses current technological challenges but also lays a solid groundwork for future innovations in the realm of computing system design.

Contents

1	Introduction	4
2	Objectives	5
3	Previous Work	6
3.1	The Concept of Network-on-Chip	6
3.2	Network-on-Chip Simulator	7
3.3	Solutions Powered By Machine Learning	8
4	Current status	10
5	Self-review	10
6	Project Plan	12
6.1	Project plan timeline	12
6.2	Expected Results and Challenges	12
7	Conclusion	13

1 Introduction

The increasing complexity of mobile applications in recent years has significantly escalated the need for advanced integration and performance in computing systems, as highlighted in various studies[1]. In this evolving landscape, the architecture of communication systems plays a critical role, profoundly influencing overall efficiency, performance, and energy consumption. Traditional connectivity methods, once the backbone of computing systems, now often fall short of meeting these demands[2]. They are commonly plagued by synchronization errors and high energy demands, as documented in recent research[3]. In response to these limitations, the innovative concept of Network-on-Chip (NoC) has emerged as a game-changing solution.

Simultaneously, the realm of Artificial Intelligence (AI), particularly Deep Neural Networks (DNNs), has witnessed remarkable progress. These advancements have catapulted AI into the forefront of various fields, such as image processing, natural language processing, and speech recognition. DNNs, characterized by their intricate multi-layer structures and high levels of accuracy, have revolutionized the way data is processed and interpreted. However, these networks also pose substantial computational challenges. They stretch the limits of traditional computing systems, complicating hardware accelerator implementations, and demanding significant resources for training and inference.

The intersection of AI and NoC technology opens up exciting possibilities. Research in this domain encompasses three main areas: utilizing AI to evaluate and predict NoC parameters, thus enhancing hardware performance; exploiting NoC's scalability to boost the execution speed and scope of AI applications; and the less-explored yet promising avenue of using AI to directly improve NoC's performance metrics, such as throughput and latency.

My research is focused on the first area, employing AI for accurate NoC parameter prediction, to optimize hardware operational efficiency. This approach not only confronts the immediate challenges inherent in AI and NoC architectures but also lays the groundwork for innovative solutions that could reshape the fields of mobile computing and AI. By developing a synergistic relationship between AI's computational power and NoC's communication efficiencies, this research aims to forge robust, efficient, and scalable systems for the future.

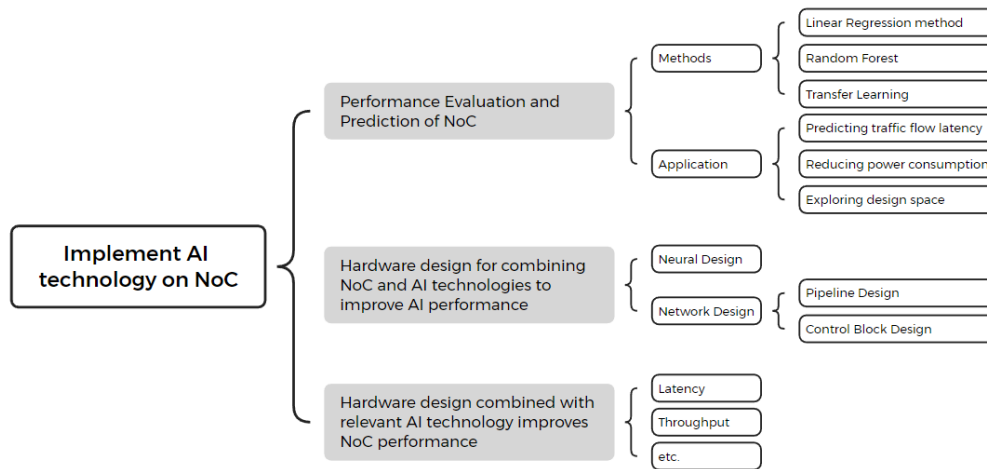


Figure 1: Classification of studies related to the combination of AI and NoC

2 Objectives

- Building and Configuring the NoC Network:** The first phase of the project is a deep exploration and application of Booksim, an open-source software, for Network-on-Chip (NoC) configuration. This involves creating a detailed 2D network structure that effectively models the NoC system. Key tasks include selecting the most suitable routing algorithms to facilitate optimal data flow, meticulously determining the size of the NoC to strike a balance between complexity and performance, and the careful process of router instantiation at each node. An integral part of this phase also involves strategically connecting channels and routers to ensure efficient communication across the network. This foundational phase is critical as it sets up the infrastructure for the entire project, aiming to establish a robust, scalable NoC framework.
- Simulation Testing with Booksim2:** This step revolves around gaining comprehensive expertise in Booksim2, a sophisticated simulation tool designed specifically for NoC networks. Mastery of this tool involves more than just understanding its functionalities; it requires manipulating the source code to independently configure various parameters, a skill critical for accurate simulation and testing. The ability to customize these parameters is essential, as it allows for a thorough exploration of the network's efficiency and highlights potential areas for improvement. This phase is crucial as it provides insights and data that inform the subsequent stages of the project.
- AI Algorithm Development for Parameter Prediction:** In the third phase, the focus shifts to AI algorithm development, starting with a detailed review and analysis of relevant literature

to identify the most suitable AI techniques for predicting NoC parameters. The initial approach involves replicating proven algorithms from academic sources, followed by attempts to enhance and optimize them. This optimization is not merely an application of existing algorithms but an exploration into improving their efficacy, thereby pushing the boundaries of AI in the context of network parameter prediction. This step is pivotal as it leverages the computational power of AI to streamline and optimize NoC operations.

- **Data Collection and Computation:** The final phase is centred on the systematic collection and computation of data. It entails learning and applying sophisticated statistical methods and data presentation techniques, as recommended in the literature. The challenge here is twofold: to accurately collect relevant data and to analyze and present it in a manner that is both insightful and comprehensible. This stage is vital as it is where the theoretical, simulation-based work is translated into tangible, measurable outcomes. The data collected and analyzed here will provide a comprehensive evaluation of the project's success and guide future research directions.

In summary, the project's objectives form a detailed and multi-dimensional roadmap for exploring the integration of AI and NoC technology. Each phase is intricately linked, with each building on the last to form a cohesive and efficient system. As the project advances, these objectives will be refined and adapted, integrating new insights and advancements in technology. This approach is designed to significantly enhance the efficiency and capabilities of NoC systems through the strategic application of AI.

3 Previous Work

3.1 The Concept of Network-on-Chip

Network-on-chip (NoC) is a communication paradigm for interconnecting various components within a chip, primarily used in system-on-chip (SoC) designs. It replaces traditional bus systems with a more scalable and efficient network-based solution. NoCs are characterized by their ability to provide high-speed data transfer between modules like processors, memory, and peripherals in a chip.

The architecture comprises several sections of wires and routers that are arranged in a grid-like manner, resembling the streets of a city. The different blocks of the city represent logical processor cores, and they are separated by wires. The clients are placed on these blocks, while the Network Interface (NI) module converts the packets generated by them[4]. Figure 2 is a visual representation of the setup.

Key features of NoCs include modularity, scalability, and flexibility, allowing them to support a wide range of applications and chip sizes[2]. They are designed to handle the increasing commu-

unication demands of modern integrated circuits, offering better performance, power efficiency, and reduced latency compared to traditional bus systems.

NoCs can be categorized into several types based on architecture: Mesh, Tree, Ring, Star, and Crossbar are common examples. Each type has its advantages and is suited for specific applications and chip layouts. For instance, mesh NoCs are popular for their simplicity and regular structure, making them suitable for large-scale integration.

In summary, NoCs represent a crucial development in chip design, addressing the limitations of traditional bus architectures and enabling more complex and efficient SoC designs.

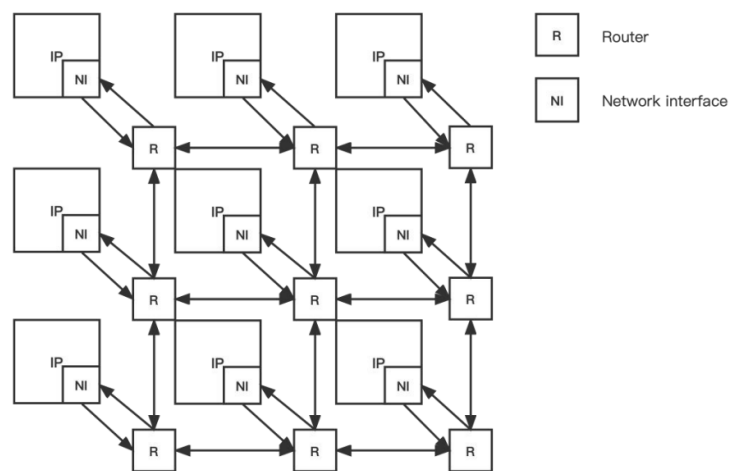


Figure 2: NoC Architecture in a Mesh Topology

3.2 Network-on-Chip Simulator

Currently, research on Network on Chip (NoC) is on the rise, and its simulation is becoming increasingly important. In this regard, there are several commonly used NoC simulation platforms available. After careful consideration, I have chosen Booksim as the NoC configuration and generation platform. The main reasons for this decision are that its resources are open, it provides more details, has a larger user base, and more.

Noxim: Noxim is an open-source NoC simulator designed for studying and designing 2D mesh Networks-on-Chip. It is widely used for its flexibility and is capable of simulating energy, performance, and thermal metrics. Its YAML-based configuration makes it user-friendly and adaptable to various research needs[5].

BookSim: Developed by Stanford University, BookSim is a cycle-accurate NoC simulator used for studying chip-scale and large-scale multiprocessor interconnection networks. It supports a range of topologies and routing algorithms and is known for its detailed simulation of network behavior[6].

NoCTweak (Tran and Baas): This SystemC-based simulator is designed for early performance exploration of NoCs, including throughput, latency, and energy estimation. It uses CMOS library cell data for post-layout timing and power estimation and supports 2D mesh topology with both synthetic and embedded traffic patterns[7].

Nirgam (Lavina Jain): A collaborative effort between the University of Southampton and Malaviya National Institute of Technology, Nirgam is an open-source, discrete event, cycle-accurate simulator supporting 2D mesh and torus topologies. It features a wormhole switching mechanism and supports source XY and OE routing mechanisms[7].

Nostrum: Created by the Nostrum Team at KTH Stockholm, this simulator is cycle-accurate and supports 2D mesh and torus topologies with wormhole, store, and forward switching mechanisms. It allows for application mapping and configuration for both best effort and guaranteed communications[7].

NOCMAP/ReliableNoC (Hu et al.): This open-source C++ mapping simulator implements BB and SA mapping algorithms. It uses the bit energy model for energy minimization and shows that BB is more efficient than SA in terms of result optimality and simulation speed[7].

3.3 Solutions Powered By Machine Learning

Integrating Artificial Intelligence (AI) with Network-on-Chip (NoC) architectures is an emerging research area that significantly impacts the design and optimization of modern computing systems. This integration is primarily explored in three domains, each focusing on different aspects of NoC and AI synergy.

AI for NoC Performance Enhancement: This domain involves leveraging AI techniques, notably deep reinforcement learning (DRL) and machine learning (ML), to enhance NoC performance. Researchers focus on optimizing NoC arbitration, routerless NoC architecture, and overall design efficiency. Significant outcomes include improved packet latency, enhanced throughput, and increased power efficiency, demonstrating AI's potential in addressing complex design challenges in NoC systems.

Biswajit Bhowmik et al. introduce a machine learning framework for NoC performance evaluation, achieving up to 94% accuracy and a 2228x speedup over traditional methods. It uses linear regression to predict metrics like latency and power consumption, offering a faster, more precise alternative[8]. Yang Li and Pingqiang Zhou presented a work based on a neural network method. This approach achieves an impressive 95% average estimation accuracy and offers a significant speedup (17.1X) for large-scale NoCs compared to BookSim2 simulations. The method significantly improves accuracy (20% to 70%) over other machine learning-based works in the field[9]. Silva, J., Kreutz, M., Pereira, M. et al. present a machine learning-based approach to predict NoC latency with high accuracy. Employing Random Forest and other tree-based classifiers, the study achieves up to 99% accuracy for audio/video applications and around 85%-90% for others, sig-

nificantly speeding up the design process for NoC-based systems[10]. Kumar A, and Talawar B. show a framework using Support Vector Regression (SVR) and Artificial Neural Networks (ANN) is proposed for NoC performance prediction. This method shows a speedup of 1500× to 2000× compared to traditional simulators[11]. Hou J, Han Q, and Radetzki M. report a 153× speedup in performance metric computation using machine learning, compared to the Noxim simulator. This speedup increases to 6844× when excluding training time. The accuracy of their predictions for fault resilience reaches up to 99.70%[12].

AI-Driven Energy Efficiency in NoC Designs: In this area, AI, especially reinforcement learning, is applied to develop energy-efficient NoC designs. The integration of AI algorithms with techniques like Power-Gating (PG) and Dynamic Voltage and Frequency Scaling (DVFS) has shown substantial improvements in reducing power consumption while maintaining system performance. This approach is crucial for sustainable computing, particularly in large-scale, high-performance computing environments.

Chen, Kun-Chih Jimmy, et al. present a significant reduction in off-chip memory accesses reported for NoC-based designs compared to conventional designs: 94.6% for LeNet, 99.6% for MobileNet, and 88.1% for VGG-16, demonstrating the efficiency of NoC-based architectures[13]. Dong, Yiping. present a novel design combining Field-Programmable Gate Array (FPGA) and Network-on-Chip (NoC) for accelerating artificial neural network (ANN) computations. This approach results in a substantial increase in computing speed, reaching over 3.1 Giga Connections Per Second (CPS)[14]. Chen, Kun-Chih, et al. show the NoC-based design significantly reduces off-chip memory accesses compared to conventional designs: 94.6% for LeNet, 99.6% for MobileNet, and 88.1% for VGG-16 showcasing NoC's efficiency in handling data-intensive DNN models[15].

AI for Efficient NoC Parameter Evaluation: This research focus is on using AI to streamline the evaluation process of NoC parameters. AI's ability to process complex data efficiently aids in accurately and rapidly assessing NoC parameters, which is essential for NoC design and optimization. The advancements in this domain are particularly relevant for addressing scalability, fault tolerance, and accurate performance prediction in various NoC configurations.

Lin, Ting-Ru, et al. demonstrate the DRL routerless design's effectiveness, showing a 3.25x increase in throughput, a 1.6x reduction in packet latency, and a 5x reduction in power. Compared to state-of-the-art routerless NoCs, it achieves 1.47x higher throughput, 1.18x reduced packet latency, 1.14x reduced average hop count, and 6.3% lower power consumption[16]. Hao Zheng and Ahmed Louri's 2019 study on an energy-efficient Network-on-Chip (NoC) design using reinforcement learning demonstrates a 26% reduction in power consumption and a 7% performance increase. Their Artificial Neural Network (ANN) design also achieves a 67% area reduction compared to traditional reinforcement learning implementations[17].

My research interests focus on combining AI with NoC to improve the efficiency of evaluating NoC parameters. This aligns with the third area, which is critical to modern computing, as the

complexity of systems requires smarter and more efficient methods of design and evaluation. By utilising AI, enabling research can make a significant contribution to the development of more efficient, robust and scalable NoC architectures. This will ultimately drive progress in the field of high-performance computing.

In summary, the confluence of AI and NoC represents a progressive shift in network design and optimization strategies, promising significant improvements in performance metrics and system architecture. This symbiosis is poised to drive innovation in network efficiency, setting a new benchmark in the field of computing system design.

4 Current status

I am currently learning how to design, configure, simulate and test NoC networks using the Booksim2 simulator. The source code structure of the simulator consists of a main folder (src) and subfolders for replaceable components and test profiles. The system is compiled using "make" to create an executable that requires a NoC configuration file as input to start the simulation.

The simulator has default parameters for use with NoC, stored in the "bookim_config.cpp" file in MAP format, and overrides the default parameters with user-configured parameters in this MAP. First, the "main()" function extracts and initialises the configuration parameters from the supplied NoC file. Then, the "Simulate()" function is called to start the simulation phase.

During the simulation, network components are instantiated and connected, and the traffic manager is initialized. It is important to understand how these parameters are set and make sure they meet specific requirements. This setup allows for a flexible configuration process that can accommodate various simulation needs and scenarios.

Booksim2 simulator's architecture uses a modular and extensible approach, which creates a comprehensive and detailed simulation environment for NoC networks. Understanding and configuring these parameters is the key to harnessing the simulator's full potential and achieving accurate, representative results for NoC network simulations.

5 Self-review

Over the past three weeks, after much deliberation and consultation with my tutor, I've refined my research direction, focusing on Network-on-Chip (NoC) and AI technology. This reorientation required re-gathering information, which initially slowed my progress. However, I have identified a platform for realizing my research on NoC, encompassing both construction and simulation testing.

During the initial phase of direction-finding, I extensively reviewed literature centred on NoC and AI technology. This exploration revealed three primary research directions, each intertwining AI and NoC differently: one for enhancing AI technology implementation, another for efficiently

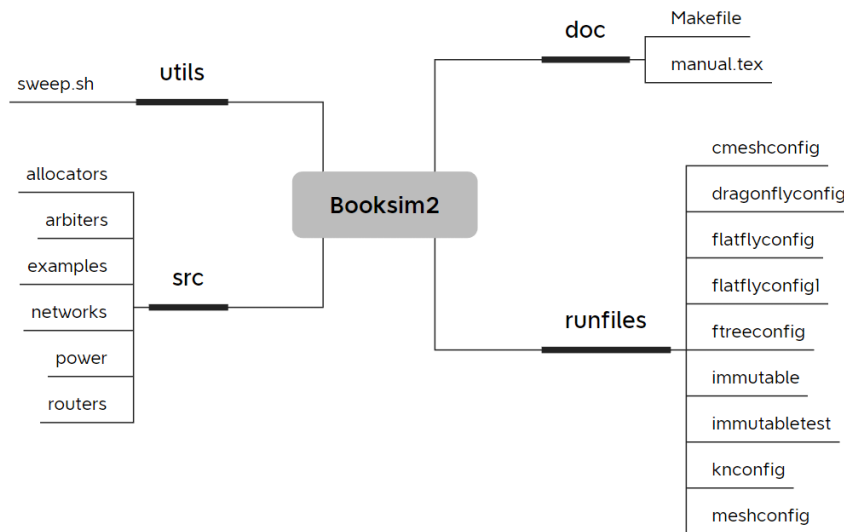


Figure 3: Booksim2 Architecture

predicting NoC parameters and the last for leveraging NoC as a platform for performance improvement. My tutor's advice influenced my final choice, the volume of existing research in these areas, and feasibility considerations.

Having pinpointed my focus, the next step involved delving into the specifics. The foundation of my research lies in NoC architecture and simulation data. To this end, I explored open-source NoC simulation platforms and, based on literature data, identified three common simulators. I selected Booksim2 for experimentation, an open-source platform favoured for its accessibility, thanks to available code analysis and tutorials by relevant experts. This platform seems user-friendly, but its efficacy and specifics will become clearer with ongoing use.

With a basic method for constructing NoC networks and a chosen simulation platform, my next step is to further narrow my literature review to one specific direction: integrating AI technology for more accurate NoC parameter prediction. The core challenge here is to develop a synergy between these two domains for enhanced parameter prediction in NoC, or application in practical scenarios. This aspect requires further exploration and research.

This phase of my study is crucial, as it involves understanding the intricacies of NoC and AI technologies and innovatively combining them to address complex computational problems. The journey ahead is challenging yet promising, with potential breakthroughs that could significantly advance the field.

6 Project Plan

6.1 Project plan timeline

As shown in Figure 4 of the timeline, the project started with a literature review focusing on NoC and AI technologies, followed by a mid-term review including resource assessment and simulation runs. The milestone progression phase includes NoC simulation, practical implementation of the AI technology, integration, and collection of experimental data. This phase will continue through the second quarter of 2024.

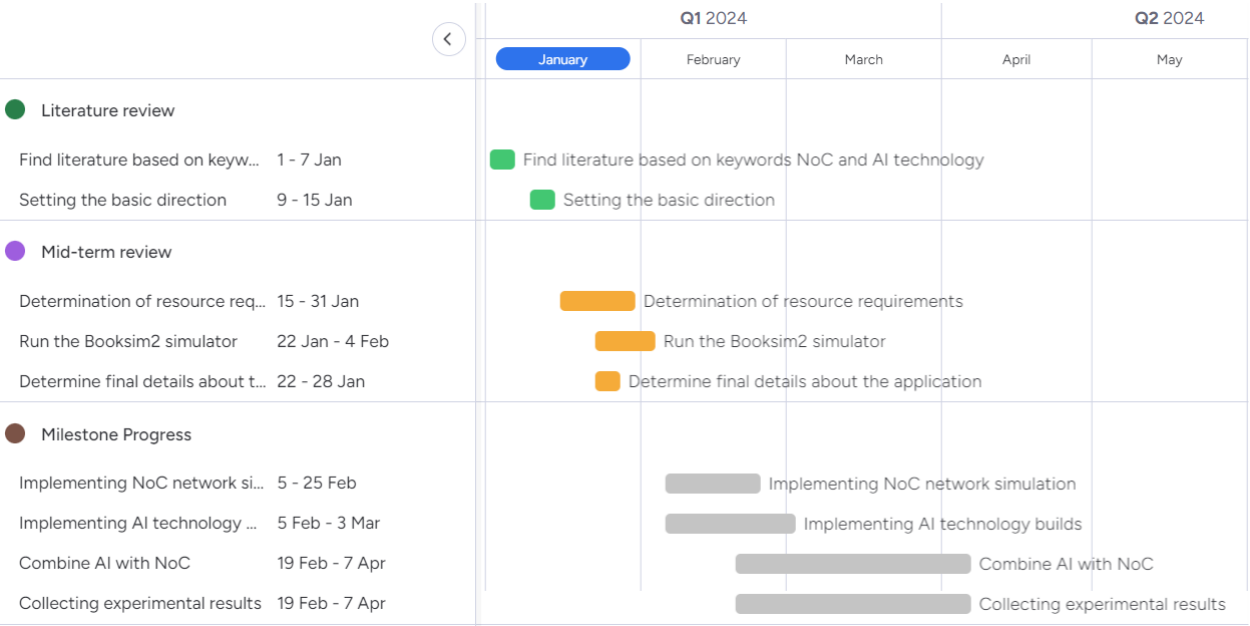


Figure 4: Gantt chart of the project plan

6.2 Expected Results and Challenges

This AI model is designed to improve the accuracy and efficiency of predicting critical parameters in NoC (Network-on-Chip) such as throughput, latency, power consumption, and area utilization. By utilizing advanced machine learning algorithms, the model will analyze historical design data, performance metrics, and simulations to provide precise and reliable predictions.

My primary objective is to evaluate and improve the use of AI in Network-on-Chip (NoC) design and analysis. This involves three critical aspects:

1. Real-world Application Performance: The main goal is to determine if implementing AI models in actual NoC applications can result in tangible performance improvements. This includes evaluating improvements in aspects such as latency, throughput, and energy efficiency in real-

world application environments. The aim is to move beyond theoretical or simulated benefits and demonstrate practical, measurable advancements in NoC operations.

2. **AI Algorithmic Enhancements:** The second aspect involves exploring the potential for further development and refinement of AI algorithms specifically for NoC analysis. This exploration seeks to understand if current AI methodologies are fully optimized or if there are unexplored avenues that could lead to more accurate predictions, faster processing times, or more efficient data handling. The focus is on both the improvement of existing algorithms and the exploration of novel AI approaches that might offer superior performance or insights.

3. **Comprehensive NoC Analysis:** The third challenge is to broaden the scope of AI applications to provide a more detailed and holistic comparison of various NoC parameters. This involves not only focusing on individual aspects such as routing efficiency or buffer size but also examining how these elements interact on an overall scale. The aim is to use AI to gain a more nuanced and comprehensive understanding of NoC design and operation, thereby enabling more informed decision-making and optimization strategies.

Addressing these challenges requires a multi-faceted approach, combining in-depth technical analysis with innovative AI applications to push the boundaries of current NoC design and performance evaluation methodologies.

7 Conclusion

This interim report encapsulates the significant strides made in the integration of Artificial Intelligence (AI) and Network-on-Chip (NoC) technologies, aiming to revolutionize the efficiency of mobile computing systems. The research has successfully outlined the potential of AI in enhancing NoC architectures, leading to more efficient, scalable, and robust computing systems. The use of Booksim and Booksim2 for NoC network configuration and simulation testing has laid a solid foundation for further exploration.

The progress of the project in developing AI algorithms for accurately predicting NoC parameters is evidence of the feasibility and impact of the interdisciplinary approach. Despite challenges in balancing complexity and performance, the project continues to push the boundaries of current computing paradigms.

Moving forward, the research aims to refine these methodologies, incorporate emerging technologies, and adapt to the evolving landscape of AI and NoC integration. The ultimate goal is to develop systems that not only meet the current demands of mobile computing but also pave the way for future innovations in this rapidly advancing field.

Bibliography

- [1] Radu Marculescu, Umit Y. Ogras, Li-Shiuan Peh, Natalie Enright Jerger, and Yatin Hoskote. Outstanding research problems in noc design: System, microarchitecture, and circuit perspectives. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 28(1):3–21, 2009.
- [2] Luca Benini and Davide Bertozzi. Network-on-chip architectures and design methods. *IEE Proceedings-Computers and Digital Techniques*, 152(2):261–272, 2005.
- [3] Jerry Zhao, Animesh Agrawal, Borivoje Nikolic, and Krste Asanović. Constellation: An open-source soc-capable noc generator. In *2022 15th IEEE/ACM International Workshop on Network on Chip Architectures (NoCArc)*, pages 1–7, 2022.
- [4] Sao-Jie Chen, An-Yeu Wu, and Jiang Xu. Networks-on-chip: Architectures, design methodologies, and case studies. *Journal of Electrical and Computer Engineering*, 2012, 04 2012.
- [5] Vincenzo Catania, Andrea Mineo, Salvatore Monteleone, Maurizio Palesi, and Davide Patti. Noxim: An open, extensible and cycle-accurate network on chip simulator. In *2015 IEEE 26th international conference on application-specific systems, architectures and processors (ASAP)*, pages 162–163. IEEE, 2015.
- [6] Nan Jiang, Daniel U. Becker, George Michelogiannakis, James Balfour, Brian Towles, D. E. Shaw, John Kim, and William J. Dally. A detailed and flexible cycle-accurate network-on-chip simulator. In *2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 86–96, 2013.
- [7] Sarzamin Khan, Sheraz Anjum, Usman Ali Gulzari, and Frank Sill Torres. Comparative analysis of network-on-chip simulation tools. *IET Computers & Digital Techniques*, 12(1):30–38, 2018.
- [8] Biswajit Bhowmik, Pallabi Hazarika, Prachi Kale, and Sajal Jain. Ai technology for noc performance evaluation. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 68(12):3483–3487, 2021.
- [9] Yang Li and Pingqiang Zhou. Fast and accurate noc latency estimation for application-specific traffics via machine learning. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2023.
- [10] Jefferson Silva, Márcio Kreutz, Monica Pereira, and Marjory Da Costa-Abreu. An investigation of latency prediction for noc-based communication architectures using machine learning techniques. *The Journal of Supercomputing*, 75:7573–7591, 2019.

-
- [11] Anil Kumar and Basavaraj Talawar. Machine learning-based framework to predict performance evaluation of on-chip networks. In *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pages 1–6. IEEE, 2018.
 - [12] Jie Hou, Qi Han, and Martin Radetzki. A machine learning enabled long-term performance evaluation framework for nocs. In *2019 IEEE 13th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoc)*, pages 164–171. IEEE, 2019.
 - [13] Kun-Chih Jimmy Chen, Masoumeh Ebrahimi, Ting-Yi Wang, Yuch-Chi Yang, and Yuan-Hao Liao. A noc-based simulator for design and evaluation of deep neural networks. *Microprocessors and Microsystems*, 77:103145, 2020.
 - [14] Yiping Dong. *A Study on Hardware Design for High Performance Artificial Neural Network by using FPGA and NoC*. PhD thesis, Waseda University, 2011.
 - [15] Kun-Chih Chen, Masoumeh Ebrahimi, Ting-Yi Wang, and Yuch-Chi Yang. Noc-based dnn accelerator: A future design paradigm. In *Proceedings of the 13th IEEE/ACM international symposium on networks-on-chip*, pages 1–8, 2019.
 - [16] Ting-Ru Lin, Drew Penney, Massoud Pedram, and Lizhong Chen. A deep reinforcement learning framework for architectural exploration: A routerless noc case study. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 99–110. IEEE, 2020.
 - [17] Hao Zheng and Ahmed Louri. An energy-efficient network-on-chip design using reinforcement learning. In *Proceedings of the 56th Annual Design Automation Conference 2019*, pages 1–6, 2019.